Linking field synopses with data collection through disease-specific networks and biobanks (HuGE networks and P3G)

Julian Little and Isabel Fortier

HuGENet Workshop, Atlanta, GA

January 24-25, 2008

How are we responding to the difficulty to identify and replicate genetic associations?

Among others, by:

- Trying to conduct studies with optimal designs
- Increasing the size of individual studies
  - Very large sample size needed
- Promoting the conduct of meta-analysis

  Data pooling limited by the heterogeneity between studies (data access rules, designs, information collected, etc.)

The Public Population Project in Genomics (P^3G) choose to support:

- **Sharing of knowledge:** Support the development of biobanks and the exchange of expertise and tools.
- **Documentation**: Provide access to descriptive information on participant organizations (design, ethical rules, measures and samples collected, etc.).
- **Harmonization**: (1) Support the development of common guidelines; (2) Identify common information that could be shared across networks of studies
- **Sharing of information:** Support the development of IT tools facilitating data exchange.

P3G Working groups:

Genomics and Biochemical Investigations

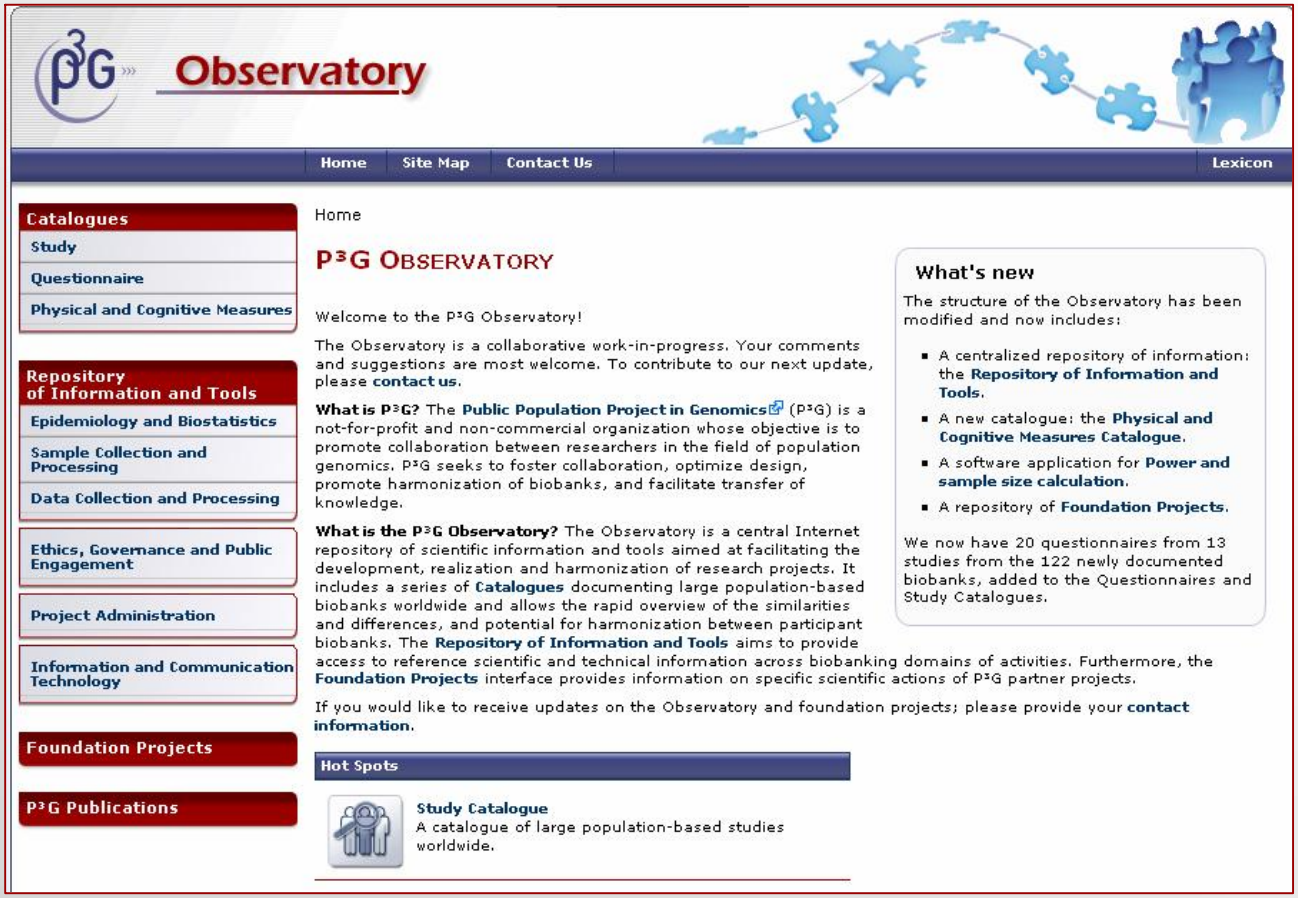
Knowledge Curation and Information Technology

Ethics, Governance and Public Engagement

Epidemiology and Biostatistics

The P³G Observatory: A collaborative work-in-progress

- Provide tools to support researchers in the harmonization, development, and realization of studies.



Observatory Catalogues

- They provide a quick and easy access to:
  - The information collected and methods in use by selected biobanks
  - Interactive tools allowing the rapid overview of the similarities and differences, and potential for harmonization between participant biobanks.

Study Catalogue

STUDY CATALOGUE

This Catalogue is a repository of standard information describing population-based studies in genomics.

To be included in the Catalogue, a study must:

- Collect or plan to collect DNA samples or biological material from which DNA can be extracted.
- Have recruited or plan to recruit more than 10 000 healthy individuals (some smaller studies are also included occasionally).

The descriptive information is collected in a two-steps process: (1) The summary information is initially gathered by our team through websites and/or publications and (2) Complete information is provided by study investigators using the P³G study description form. For studies with only a summary description, the scope and quality of information collected is limited, and the search form provides fewer data. If you would like your study to be included in the catalogue, please complete the **study description form** and send it back to us by **email**.

Catalogue Current Content	
Number of Studies	122
With Summary Information Only	81
With Complete Information	41

Comparison Tools

Three tables are provided in xls format to compare information on studies.

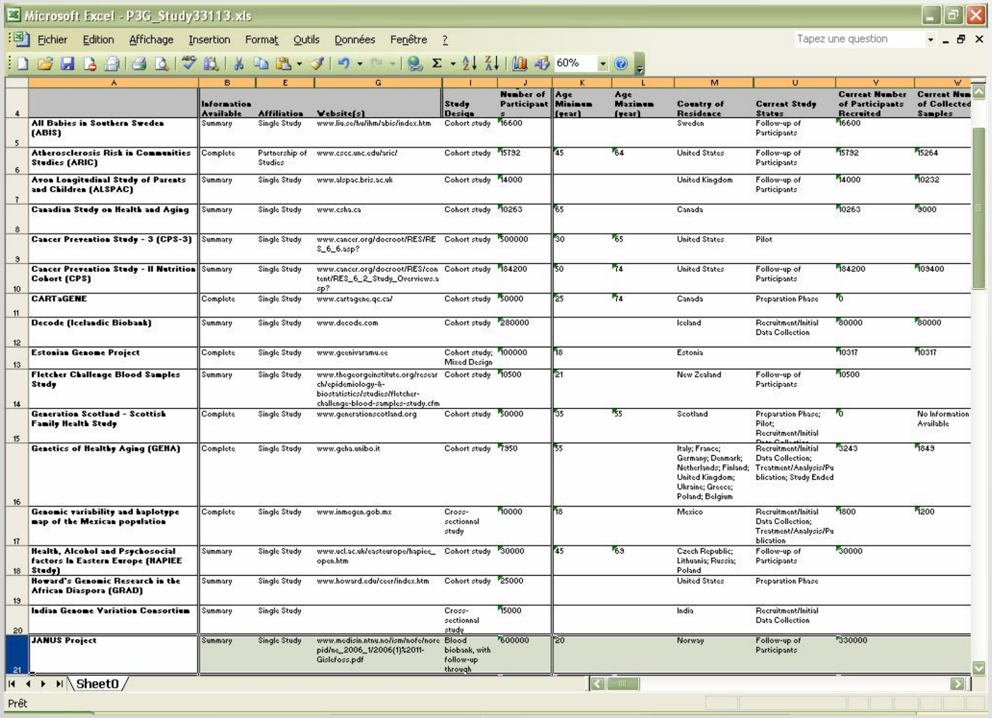
Table	Description
Summary Information	Summary information.
Principal Topics of Interest	Principal information collected.
Ethics & Governance	Information on ethics and governance.

122 large population-based biobanks

Study Catalogue

Description of the population-based biobanks (contact information, website, objectives, design, selection criteria, current status, etc.)

Comparison Tables (Summary information; principal information collected; ethics and governance issues)



Microsoft Excel - P3G_Study33113.xls

	Information Available	Affiliation	Website(s)	Study Design	Number of Participants	Age Minimum (year)	Age Maximum (year)	Country of Residence	Current Study Status	Current Number of Participants Recruited	Current Number of Collected Samples
All Babies in Southern Sweden (ABIS)	Summary	Single Study	www.liu.se/hu/ihm/abis/index.htm	Cohort study	16600			Sweden	Follow-up of Participants	16600	
Atherosclerosis Risk in Communities Studies (ARIC)	Complete	Partnership of Studies	www.cscc.unc.edu/aric/	Cohort study	15792	45	64	United States	Follow-up of Participants	15792	15264
Avon Longitudinal Study of Parents and Children (ALSPAC)	Summary	Single Study	www.alspac.bris.ac.uk	Cohort study	14000			United Kingdom	Follow-up of Participants	14000	10232
Canadian Study on Health and Aging	Summary	Single Study	www.csha.ca	Cohort study	10263	65		Canada		10263	9000
Cancer Prevention Study - 3 (CPS-3)	Summary	Single Study	www.cancer.org/docroot/RES/RES_6_6.asp?	Cohort study	500000	30	65	United States	Pilot		
Cancer Prevention Study - II Nutrition Cohort (CPS)	Summary	Single Study	www.cancer.org/docroot/RES/content/RES_6_2_Study_Overviews.asp?	Cohort study	184200	50	74	United States	Follow-up of Participants	184200	109400
CARTaGENE	Complete	Single Study	www.cartagene.qc.ca/	Cohort study	50000	25	74	Canada	Preparation Phase	0	
Decode (Icelandic Biobank)	Summary	Single Study	www.decode.com	Cohort study	280000			Iceland	Recruitment/Initial Data Collection	80000	80000
Estonian Genome Project	Complete	Single Study	www.geenivaramu.ee	Cohort study; Mixed Design	100000	18		Estonia		10317	10317
Fletcher Challenge Blood Samples Study	Summary	Single Study	www.thegeorgeinstitute.org/research/epidemiology-&-biostatistics/studies/fletcher-challenge-blood-samples-study.cfm	Cohort study	10500	21		New Zealand	Follow-up of Participants	10500	
Generation Scotland - Scottish Family Health Study	Complete	Single Study	www.generationscotland.org	Cohort study	50000	35	55	Scotland	Preparation Phase; Pilot; Recruitment/Initial Data Collection	0	No Information Available
Genetics of Healthy Aging (GEHA)	Complete	Single Study	www.geha.unibo.it	Cohort study	7950	55		Italy; France; Germany; Denmark; Netherlands; Finland; United Kingdom; Ukraine; Greece; Poland; Belgium	Recruitment/Initial Data Collection; Treatment/Analysis/Publication; Study Ended	3243	1849
Genomic variability and haplotype map of the Mexican population	Complete	Single Study	www.inmegen.gob.mx	Cross-sectional study	10000	18		Mexico	Recruitment/Initial Data Collection; Treatment/Analysis/Publication	1800	1200
Health, Alcohol and Psychosocial factors In Eastern Europe (HAPIEE Study)	Summary	Single Study	www.ucl.ac.uk/easteurope/hapiee_open.htm	Cohort study	30000	45	69	Czech Republic; Lithuania; Russia; Poland	Follow-up of Participants	30000	
Howard's Genomic Research in the African Diaspora (GRAD)	Summary	Single Study	www.howard.edu/ceer/index.htm	Cohort study	25000			United States	Preparation Phase		
Indian Genome Variation Consortium	Summary	Single Study		Cross-sectional study	15000			India	Recruitment/Initial Data Collection		
JANUS Project	Summary	Single Study	www.medisin.ntnu.no/ism/nofe/norepid/ne_2006_1/2006(1)%2011-Gislefoss.pdf	Blood biobank, with follow-up through	600000	20		Norway	Follow-up of Participants	330000	

Sheet0

Prêt

Number of participants targeted (recruited or to be recruited) (N=105)

Number of participants	Number of studies	Number of participants TARGETED
Less than 49 000	60	1,100,000
50 000 to 99 000	15	1,000,000
100 000 to 499 000	23	3,600,000
500 000 and more	8	5,600,000
		Total: 11,300,000

Coming soon…

- A catalogue of networks will be added.
  - Studies part of a network will be described using the standard description format
  - Will include different types of networks (disease-oriented networks could be included)
  - Will be develop in collaboration and will use a common interface

Questionnaire Catalogue

CROSS-SECTIONAL QUESTIONNAIRE CATALOGUE

The Questionnaire Catalogue includes reference questionnaires from large population-based studies and aims to facilitate the development and harmonization of questionnaires. The catalogue gives access to the cross-sectional questionnaires, their methods of administration, and provides an overview of the similarities and differences of the information collected by participant studies. References to other questionnaires of interest are also available in the **Repository of Reference Questionnaires.**

Through a tree of key-words, a **Search Tool** gives quick and easy access to the information collected by a given study and allows comparison between studies. A **Comparison Table** of the information collected by the studies can be viewed via the web browser or can be downloaded as a table in **Excel format.**

Catalogue Current Content	
Number of Studies	13
Number of Questionnaires	20

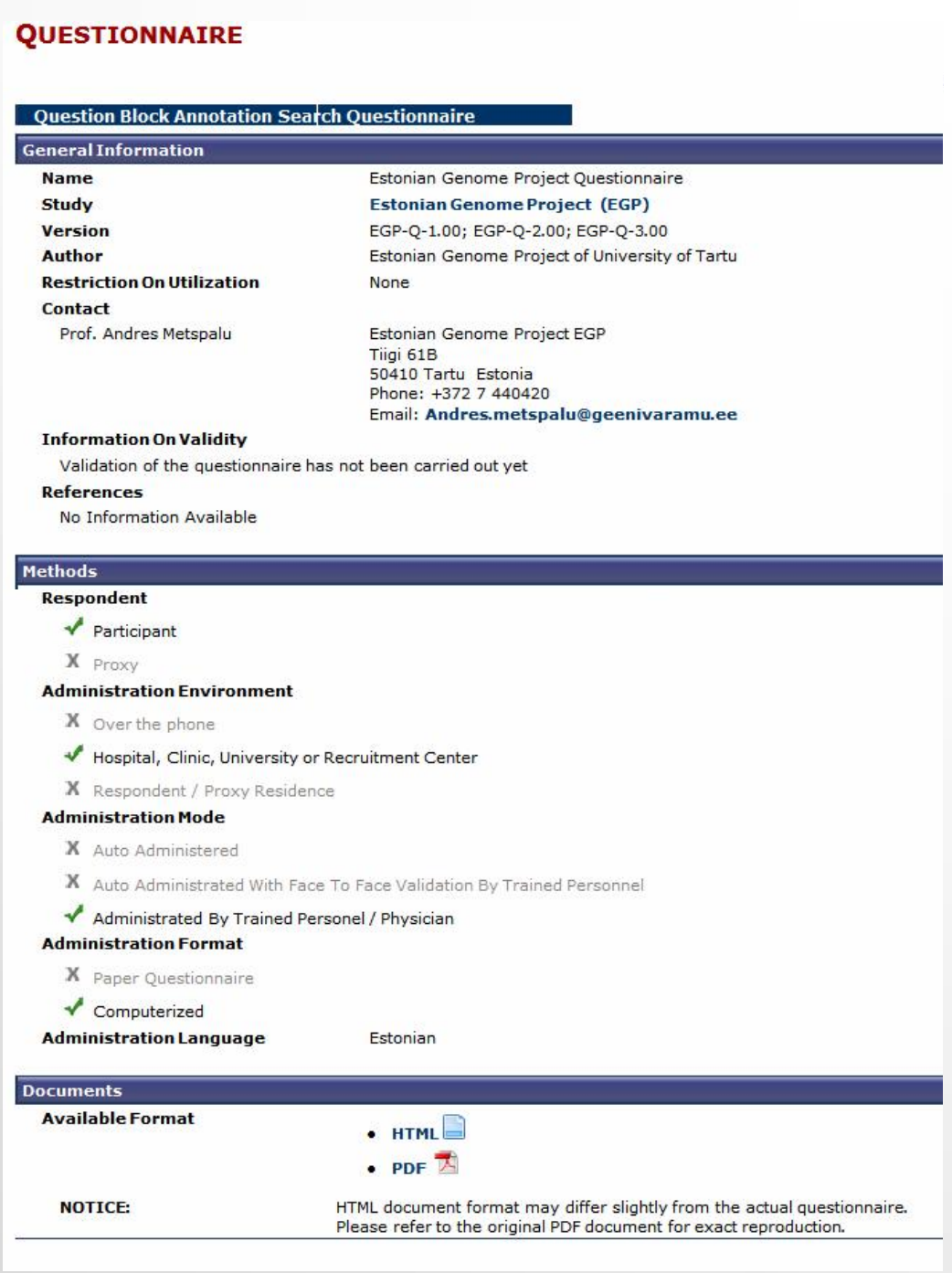
13 studies;
20 questionnaires

List of Questionnaires

1. **1998 Health and Social Survey (HSS)**
   - 1998 Health and Social Survey; Lifestyle and Health Questionna
2. **Cancer Prevention Study - 3 (CPS-3)**
   - CPS Questionnaire for Men
   - CPS Questionnaire for Women
3. **Estonian Genome Project (EGP)**
   - Estonian Genome Project Questionnaire
4. **European Prospective Investigation into Cancer and Nutrition (EPIC)**
   - 24 Hour Diet Recall

Questionnaire Catalogue

General information, methods and questionnaires



QUESTIONNAIRE

Question Block Annotation Search Questionnaire

General Information

Name	Estonian Genome Project Questionnaire
Study	**Estonian Genome Project (EGP)**
Version	EGP-Q-1.00; EGP-Q-2.00; EGP-Q-3.00
Author	Estonian Genome Project of University of Tartu
Restriction On Utilization	None
Contact	
Prof. Andres Metspalu	Estonian Genome Project EGP Tiigi 61B 50410 Tartu Estonia Phone: +372 7 440420 Email: **Andres.metspalu@geenivaramu.ee**

Information On Validity

Validation of the questionnaire has not been carried out yet

References

No Information Available

Methods

Respondent

- ✓ Participant
- X Proxy

Administration Environment

- X Over the phone
- ✓ Hospital, Clinic, University or Recruitment Center
- X Respondent / Proxy Residence

Administration Mode

- X Auto Administered
- X Auto Administrated With Face To Face Validation By Trained Personnel
- ✓ Administrated By Trained Personel / Physician

Administration Format

- X Paper Questionnaire
- ✓ Computerized

Administration Language Estonian

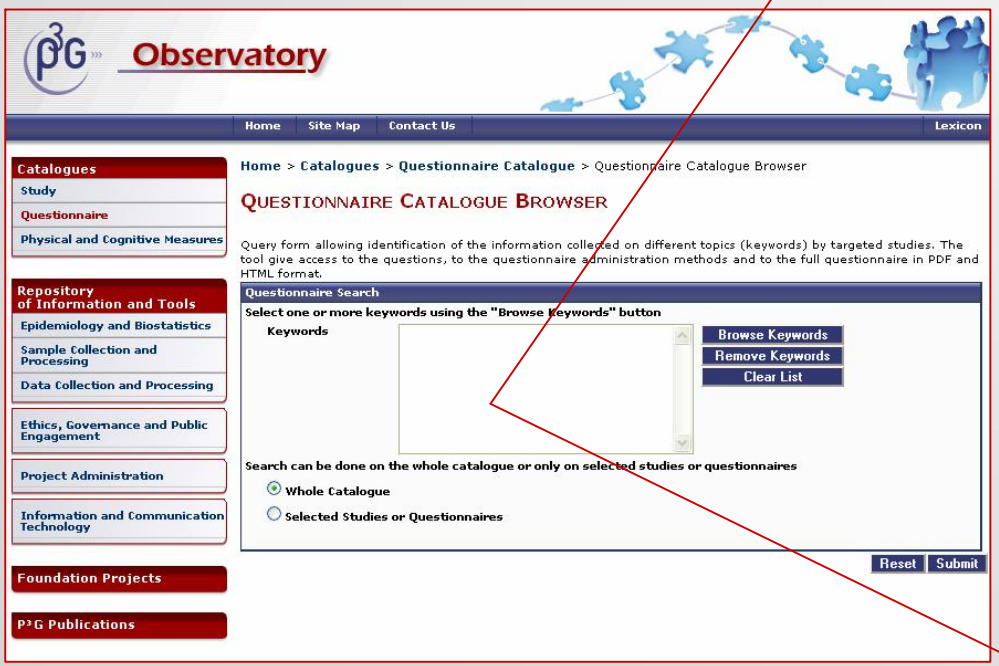
Documents

Available Format

- HTML
- PDF

NOTICE: HTML document format may differ slightly from the actual questionnaire. Please refer to the original PDF document for exact reproduction.

Questionnaire content: Search by keywords



- Health Information
  - Diseases history (ICD10)
  - Health assessment
  - Medical care system/procedures
  - Medication intake (ATC/ADD)
  - Women's health
  - Men's health
  - Early life conditions
  - Perception of health/quality of life
  - Familial diseases history
- Physical Conditions
  - Dermatology
  - Anthropometric information
  - Handedness
- Life Habits/Behaviours
  - Smoking/Tobacco use
  - Alcohol intake
  - Drug intake
  - Nutrition
  - Physical activity
  - Weight related behaviours
  - Sleep
  - Sexual behaviours
  - Travels
  - Early life/childhood
- Sociodemographic Characteristics
  - **Birth location**
    - Subjects birth location
    - Subjects family birth location
  - Citizenship/residency status
  - Ethnicity/Race
  - Language
  - Marital status
  - Education level
  - Income
  - Working status
  - Religion
- Physical Environment
  - Exposures at work
  - Environmental exposures
  - Early life and in utero exposures
- Social Environment
  - Familial and social environment
  - Working social environment

Comparison table

Life Habits/Behaviours	BNADN	STR	HSS	KORA-gen	MCPMR	KSCDC	TTP	NHANES III	EGP	EPIC	CPS-3	GS-SFHS
Smoking/tobacco use	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Alcohol use	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓
Illicit drug use		✓	✓						✓			
Nutrition	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
Food intake list	✓	✓		✓		✓	✓	✓	✓	✓	✓	✓
Intake of milk products	✓	✓		✓		✓	✓	✓	✓	✓	✓	✓
Intake of meat, eggs, fish and alternatives	✓	✓		✓		✓	✓		✓	✓	✓	✓
Intake of vegetables and fruits	✓	✓		✓		✓	✓	✓	✓	✓	✓	✓
Intake of cereals, bread, starches	✓	✓		✓		✓	✓		✓	✓	✓	✓
Intake of sweets, baked goods	✓	✓		✓			✓		✓	✓	✓	
Intake of fat		✓		✓			✓			✓	✓	
Intake of water				✓							✓	
Intake of beverages (other than water)	✓	✓		✓		✓	✓		✓	✓	✓	
Intake of salt and food supplements		✓				✓	✓	✓	✓	✓	✓	✓
Perception of nutritional habits		✓	✓							✓		

Questionnaire catalogue: Bloc of questions

QUESTION BLOCK

General Information

Block ID	100073_11
Study	National DNA Bank - BancoADN (BNADN)
Questionnaire Name	Spanish DNA Bank Questionnaire
Keywords	• Alcohol use

Question(s)

NOTICE: In spite of our best efforts, the formating of the text below may differ slightly from the actual questionnaire. Please ref
original. **PDF document** for exact reproduction.

ALCOHOL

32. How much alcohol do you habitually consume?

Wine:
Do you drink wine with your meals?
|___| Yes, how many glasses (10cl)?
|___| No
Do you drink wine outside mealtimes?
|___| Yes, how many glasses (10cl)?
|___| No

Beer:
Do you drink beer every day?
|___| Yes, how many beers (33cl)?
|___| No
Do you drink beer at the weekends?
|___| Yes, how many beers (33cl)?
|___| No

Spirits:
Do you drink spirits every day?

Physical and Cognitive Measures Catalogue

PHYSICAL AND COGNITIVE MEASURES CATALOGUE

IMPORTANT! This catalogue is currently under piloting. For any comments or suggestions, please **contact us**.

Under Piloting

The catalogue provides an overview of the physical and cognitive measures collected by participant biobanks. When available, it also presents the standard operating procedure in place. The **Comparison Table** allows comparison of the similarities and differences of the measures collected among studies through a tree of key-words, developed upon the WHO International Classification of Functioning, Disability and Health (ICF).

Catalogue Current Content

Number of Studies 6

6 studies

List of Participating Studies

1. **KORA-Life and health in the Augsburg Region (KORA-gen)**
2. **Kadoorie Study of Chronic Disease in China (KSCDC)**
3. **Marshfield Clinic Personalized Medicine Research Project (MCPMR)**
4. **National DNA Bank - BancoADN (BNADN)**
5. **National Health and Nutrition Examination Survey III (NHANES III)**
6. **The Tomorrow Project (TTP)**

Physical and Cognitive Measures Catalogue

- **General information**
- **List of measures collected**
  - Body structures
  - Body functions: based on the International Classification of Functioning, Disability and Health (ICF-WHO)
- **Administration mode**:
- **Standard operating procedures (if available)**

PHYSICAL AND COGNITIVE MEASURES KEYWORDS

Measures keywords	Studies					
BODY STRUCTURES	BNADN	KSCDC	KORA-gen	MCPMR	NHANES	TTP
ANTHROPOMETRIC STRUCTURES	✓	✓	✓	✓	✓	✓
Anthropometric measures	✓	✓	✓	✓	✓	✓
Weight	✓	✓	✓	✓	✓	✓
Standing height	✓	✓	✓	✓	✓	✓
Sitting height		✓			✓	
Waist circumference		✓	✓		✓	✓
Hip circumference		✓	✓			
Other anthropometric measures					✓	
Body composition		✓		✓	✓	
Bioimpedance		✓			✓	
Bone density				✓	✓	
Skin fold thickness					✓	
Other body composition measures					✓	
STRUCTURES OF THE CARDIOVASCULAR, IMMUNOLOGICAL AND RESPIRATORY SYSTEMS*			✓			

BODY FUNCTIONS*	BNADN	KSCDC	KORA-gen	MCPMR	NHANES	TTP
MENTAL FUNCTIONS*		✓		✓		
Global mental functions (b110-b139)*		✓		✓		
Intelligence						
Congitives functions		✓		✓		
Other global mental function measures						
Specific mental functions (b140-b189)*					✓	
Attention						
Memory						
Handedness					✓	
Psychomotor function					✓	
Other specific mental function measures						
SENSORY FUNCTIONS AND PAIN*				✓	✓	
Seeing and related functions (b210-b229)*				✓	✓	
Vision				✓	✓	
Hearing and vestibular functions (b230-b249)*					✓	
Hearing					✓	
FUNCTIONS OF THE CARDIOVASCULAR, HAEMATOLOGICAL, IMMUNOLOGICAL AND RESPIRATORY SYSTEMS*	✓	✓	✓	✓	✓	
Functions of the cardiovascular system (b410-b429)*	✓	✓	✓	✓	✓	
Blood pressure	✓	✓	✓	✓	✓	
Heart rate (at rest)	✓	✓	✓	✓		
Heart rate (under stress)						

Next steps:

- **Next catalogues update, May 2008**
  - Studies
  - **Networks**
  - Questionnaires
  - Physical and cognitive measures
  - **Ethics and governance**
  - **DNA processing and storage**

- **Future developments: Addition of new catalogues and inclusion of the longitudinal information collected**

Repository of Information and Tools

Provides access to reference information (documents, websites, etc.) in different domains of biobanking:

Repository of Information and Tools

Structure:
Bioscience section

Conceptualization
Design and Conduct
Analysis
Dissemination

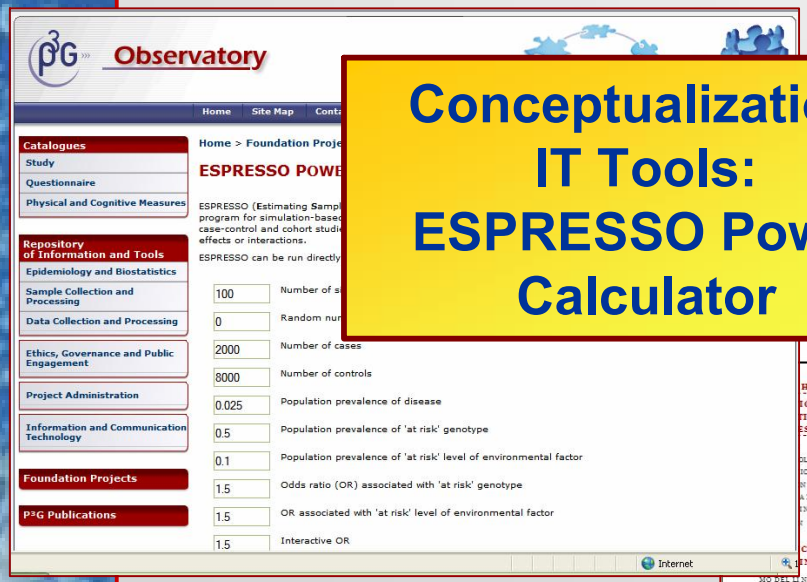
In Theory (guidelines…)
In Practice (SOPs…)
IT Tools

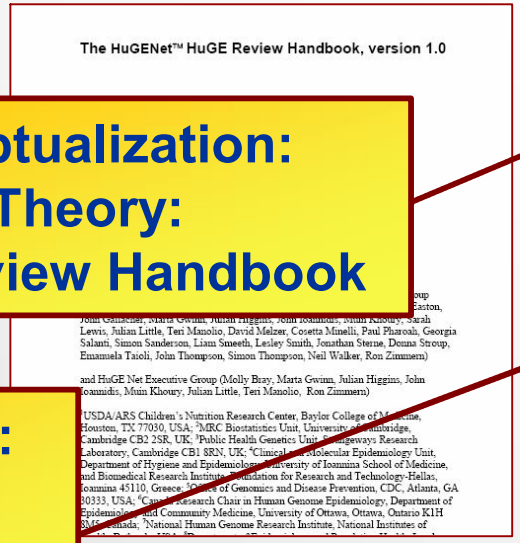
Epidemiology and Biostatistics Examples:

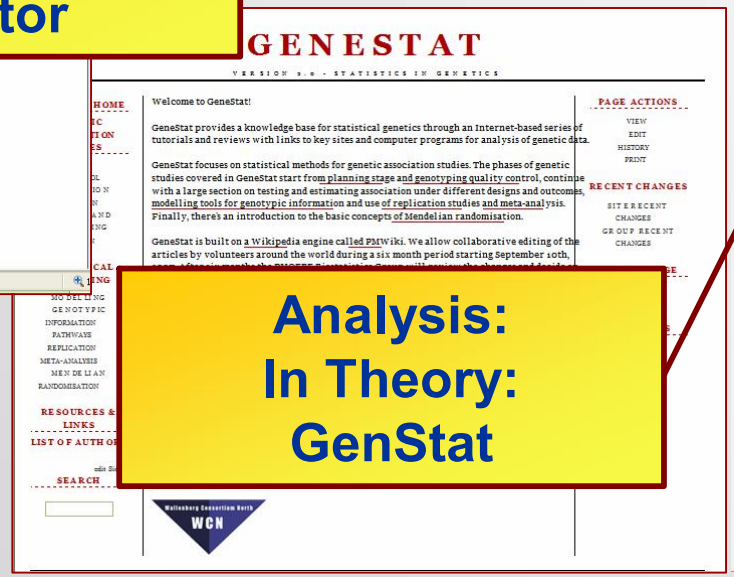
Conceptualization:
In Theory:
HuGE Review Handbook

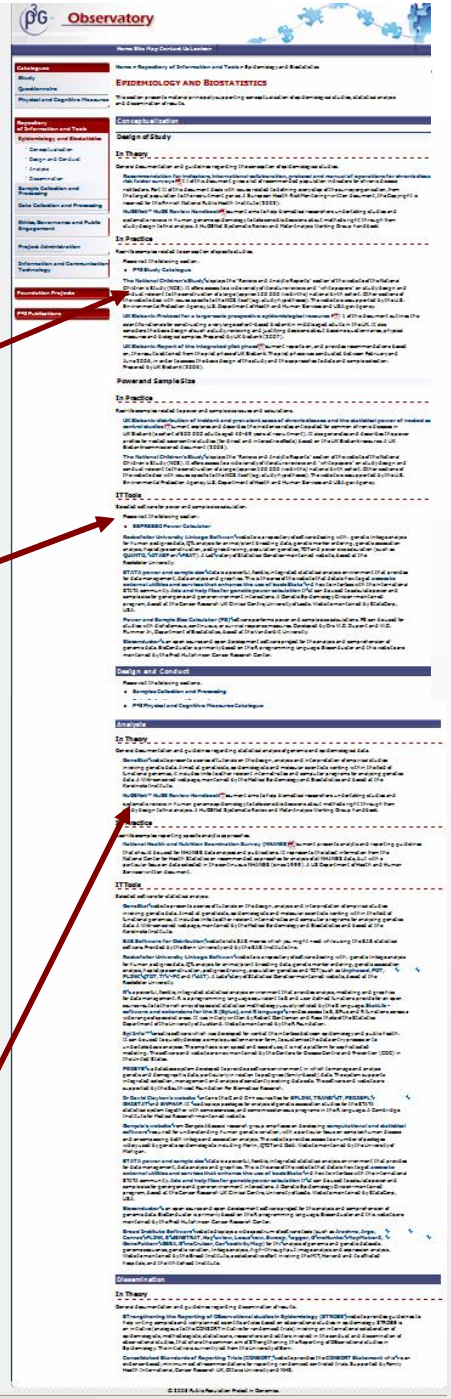
Conceptualization:
IT Tools:
ESPRESSO Power Calculator

Analysis:
In Theory:
GenStat









Sample Collection and Processing Examples:

Design and Conduct, In Theory:
Best Practices for Repositories I: Collection, Storage, and Retrieval of Human Biological Materials for Research (ISBER)


CELL PRESERVATION TECHNOLOGY
Volume 3, Number 1, 2005
© Mary Ann Liebert, Inc.

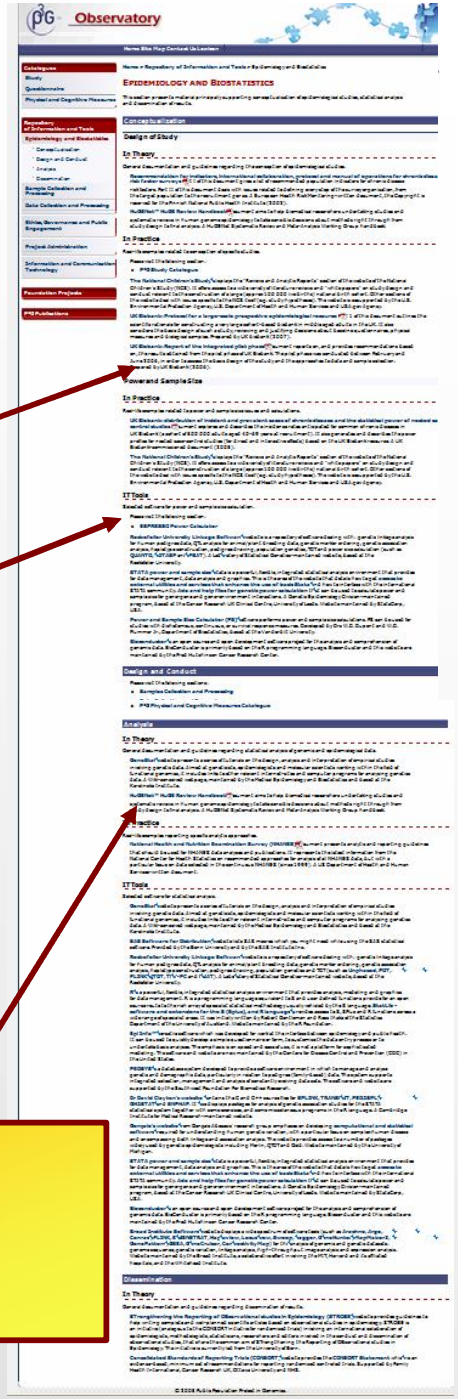
members to provide repository professionals with a comprehensive foundation for the guidance of repository activities. These Practices reflect input from individuals within and outside ad hoc committees. Best Practices will be reviewed periodically and will be revised to reflect advances in research and technology. All revisions are subject to approval by the ISBER Council. These Practices reflect the most effective approaches to the establishment and running of specimen collection facilities and are not intended as required practices.

The focus of this first version of the ISBER Best Practices is on the management of human specimen collections. It is ISBER's intention to broaden the focus to include Best Practices surrounding other specimen types in subsequent versions. Likewise, because most of the initial contributions to this edition were from individuals based in the United States, the Practices described here primarily reflect U.S. perspectives. ISBER plans to broaden the scope of the Practices in future editions to include those from other nations in order to make this document more representative of international perspectives and experience.

Best Practices are reviewed periodically and revised to incorporate improved application and research findings that would affect repository work. The reader is advised to check the ISBER website (http://www.isber.org) to be sure that the most recent version is available for use.

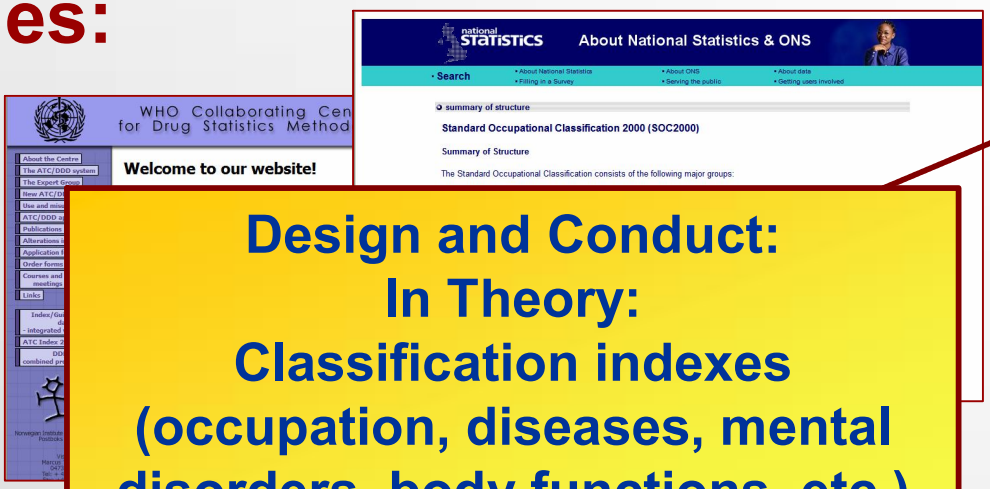
Editor-in-Chief: Karen E. Pitt, Ph.D.

Design and Conduct In Practice:
Bancoadn: Spanish blood-based National DNA Bank

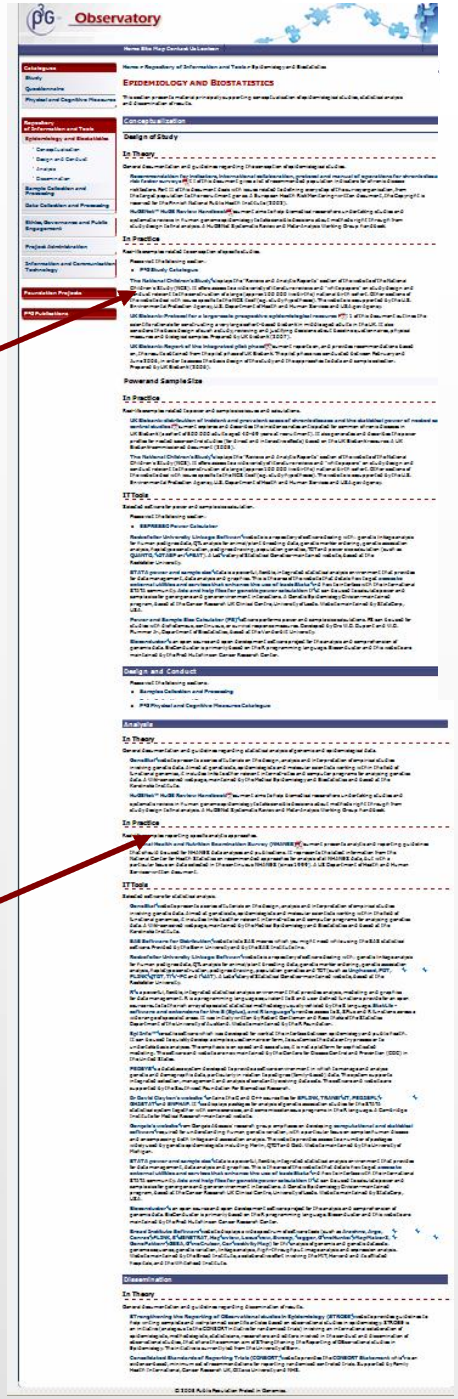


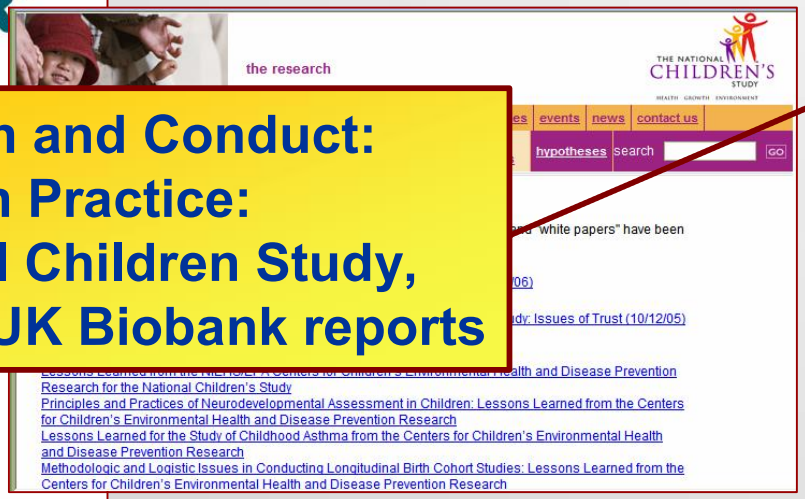
Analysis: IT Tools SeattleSNPs

Data collection and processing Examples:



Design and Conduct:
In Theory:
Classification indexes (occupation, diseases, mental disorders, body functions, etc.)





Design and Conduct:
In Practice:
National Children Study, NHANES, UK Biobank reports

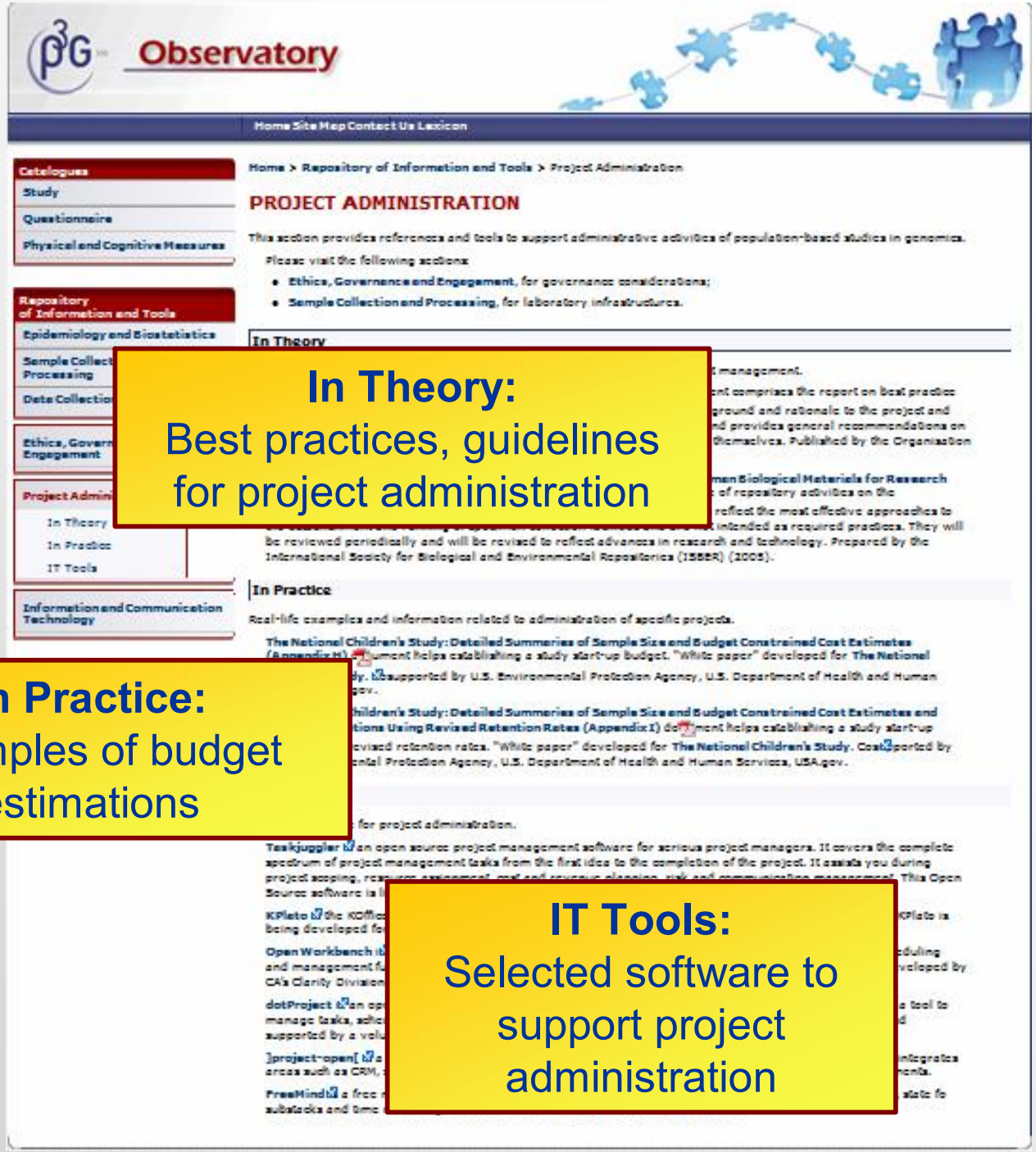
Ethics, Governance and Public Engagement

In Theory:
Legislations specific to biobanks
Selected Guidelines
Selected opinion, background documents, etc.

In Practice:
Application of theory/legislation to a given biobank

Tools:
PopGen, HUMGEN

Project Administration



Information and Communication Technology

Example:
Obiba Website
Open source management system for biobanks

INFORMATION AND COMMUNICATION TECHNOLOGY

This section lists the software referred to in the other sections of the Repository.

	Biosciences			
	Epidemiology and biostatistics	Sample collection and processing	Data collection and processing	Project Administration
American National Center for Biotechnology Information (NCBI)		✓		
Bioconductor	✓	✓		
Broad Institute Software Including Arachne, Argo, Conrad, PLINK, EIGENSTRAT, Haploview, Locusview, Sweep, Tagger, GeneHunter, HapMaker3, GenePattern, GSEA, GeneCruiser, ConnectivityMap	✓	✓		
CLC Free Workbench		✓		
dotProject				✓
Dr David Clayton's website Including SPLINK, TRANSMIT, PED2SPL, GM2STAT and SNPHAP	✓	✓		
Dr David Reich's Laboratory Including ANCESTRYMAP, EIGENSOFT and EIGENSTRAT.		✓		
Epi Info	✓		✓	
ESPRESSO Power Calculator	✓			
European Bioinformatics Institute (EBI)		✓		
FreeMind				✓
GeneSNPs		✓		
GeneStat	✓	✓		
Genetic Association Database (GAD)		✓		
Genetic Power Calculator		✓		
Gonçalo's Software	✓	✓		
HapMap Including genome browser and the Encode Project.		✓		
KPilot				✓
Matthews Stephens's website Including BIMBAM, fastPHASE, HOTSPOTTER, PHASE and SCAT		✓		
The Obiba Project		✓		
Open Workbench				✓
PEDSYS	✓	✓	✓	

Foundation Projects

IWG1

- DNA Quantity and Quality Control (Q2C)
- Harmonization and Quality Control of Lipidomics Analysis

IWG2

- The Obiba Project

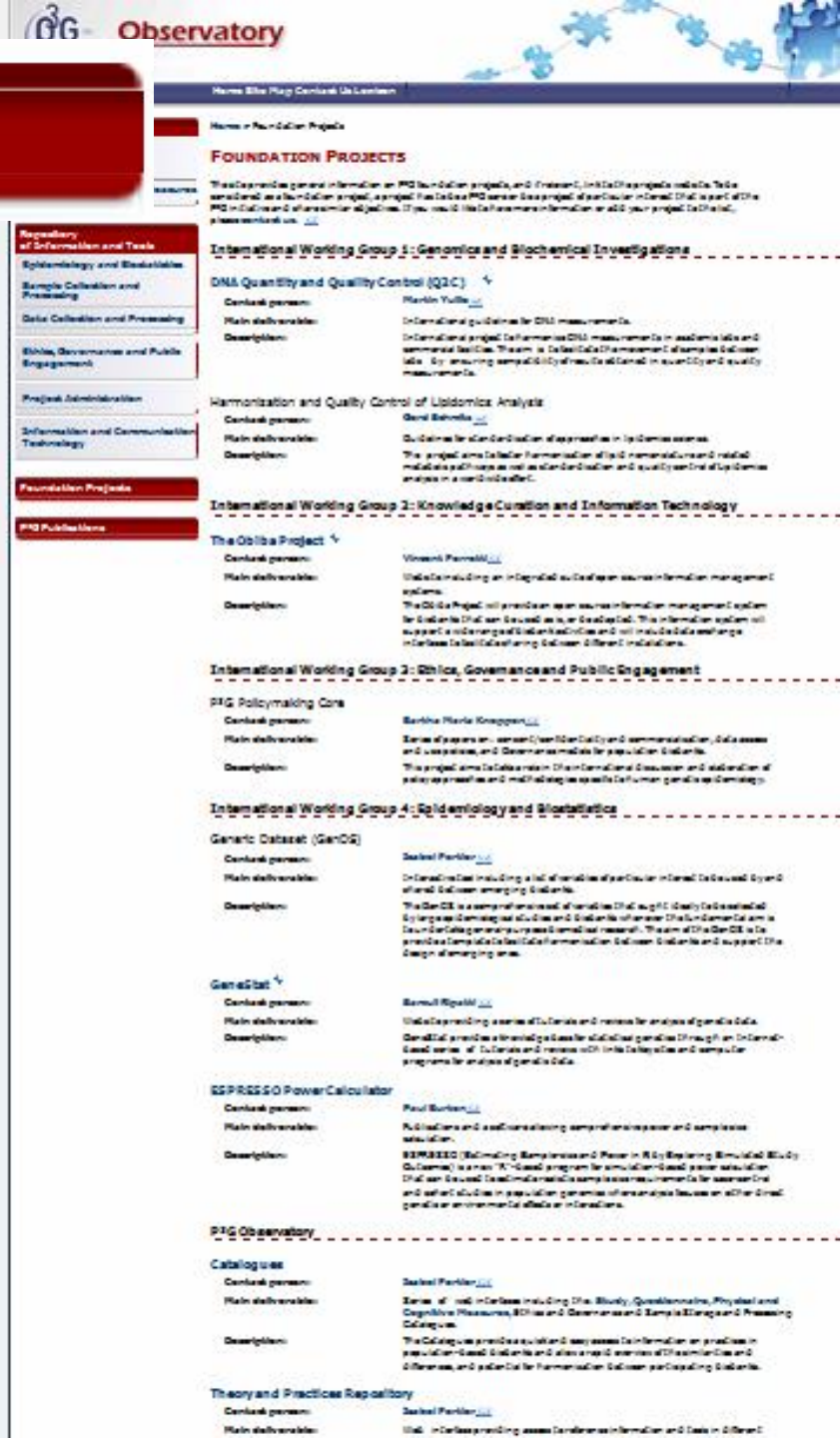
IWG3

- P3G Policymaking Core

IWG4

- Generic Dataset
- GeneStat
- ESPRESSO Power Calculator

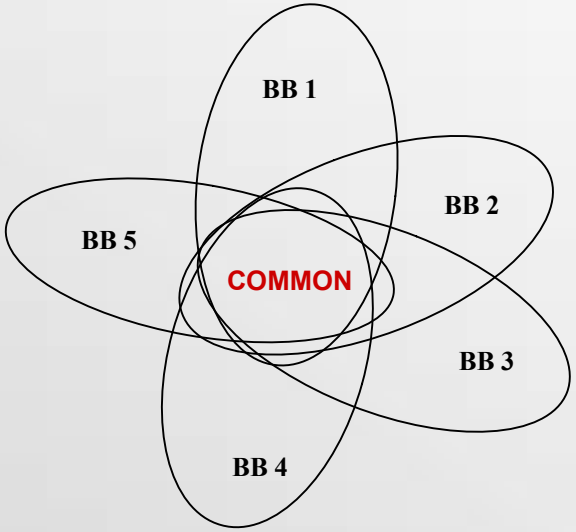
P3G Observatory



One example:
The Reference Datasets

Aim:

- Provide a template to facilitate harmonization between biobanks and support the design of emerging ones.



Characteristics of the Generic - Reference Dataset :

- Set of variables that is comprehensive enough to ensure the realization of valid research
- Simple enough to allow implementation in a variety contexts
- Small enough to encourage buy-in.
  - **NOT** a prescriptive list of **all** the variables to be collected by a biobank. Approximately 250 variables.
- The first initiative focused on middle-aged subjects and information collected at baseline
- Complementary to development of specialized datasets for particular interests (*e.g.* particular diseases, environmental exposure, etc.).

Some examples of domains covered

Health outcomes

- Cancer; diabetes; stroke; myocardial infarction; familial history of cancer, etc. ;

Health determinants

- Smoking, alcohol intake, birth location (subjects, parents and grand-parents), education, income, passive smoking exposure, working status, physical activity

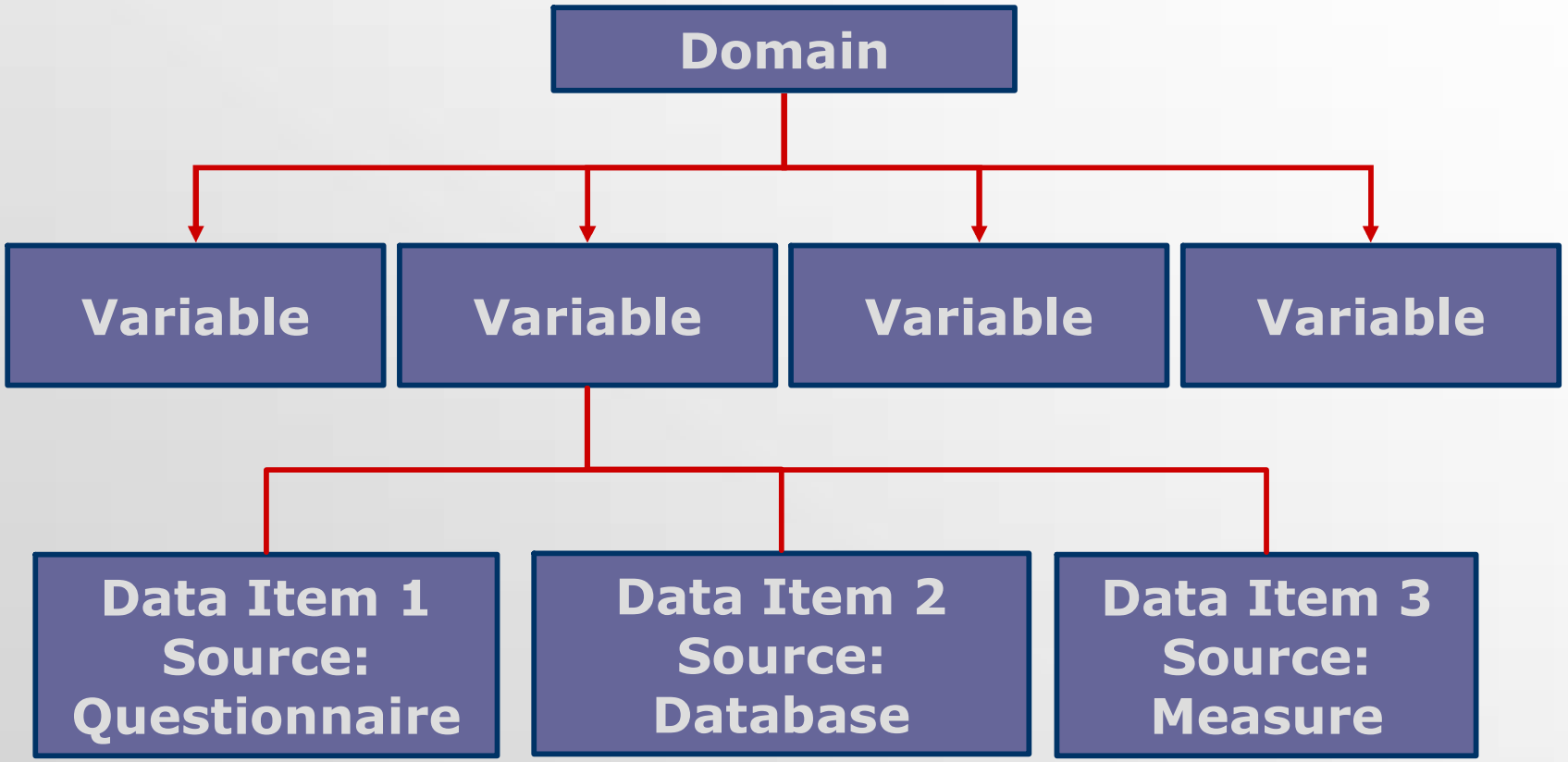
Physical measures

- Anthropometric measures, resting heart rate, blood pressure

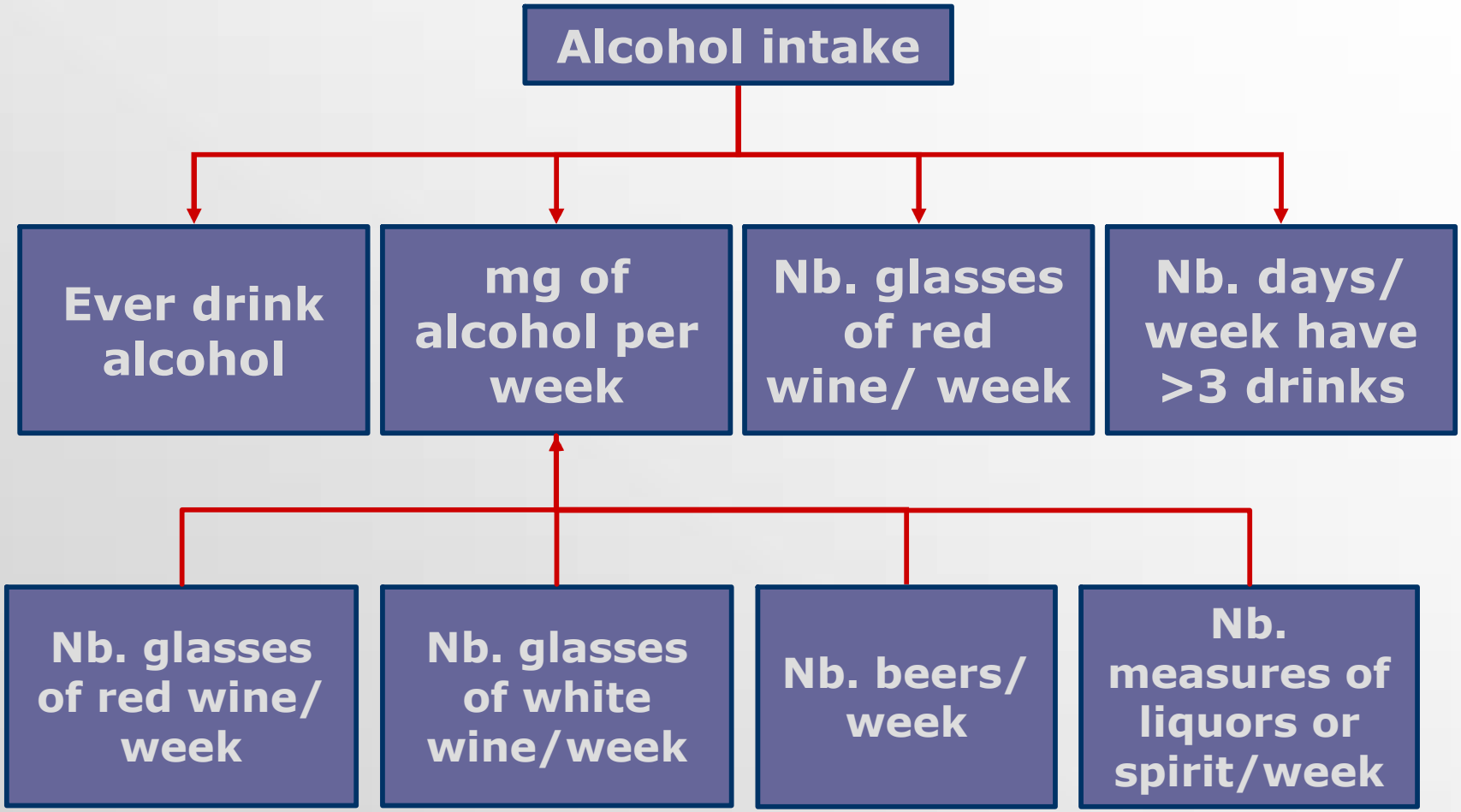
Current status:

- Initial version (January 2007)
- Piloted by 4 biobanks
  - Lifelines (Netherlands), Joondalup Family Health Study (Australia), CARTaGENE (Canada), Cardiovascular survey (Canada)
- Updated version to be finalized: March 2008
- Web interface under development: Online May 2008
- New reference datasets (Cancer, etc.) to be developed

Conceptual Framework



Example


Alcohol intake
Ever drink alcohol
mg of alcohol per week
Nb. glasses of red wine/ week
Nb. days/ week have >3 drinks
Nb. glasses of red wine/ week
Nb. glasses of white wine/week
Nb. beers/ week
Nb. measures of liquors or spirit/week

Coming soon…

- Online interactive tool available
- The web interface will provide access to:
  - *Technical information*:
    - Definitions and formats
    - Link to relevant ontologies
  - *Scientific information:*
    - Relevance
    - References
    - Coverage among participant biobanks (%)
    - Link to potential questions (or SOP)
    - Etc.
- Could host other reference datasets (disease-specific, etc.)

Domain

Variables

Data Items